How many square occurrences must a binary sequence contain?

Gregory Kucherov* Pascal Ochem† Michaël Rao‡

Submitted: Dec 3, 2002; Accepted: Dec 15, 2002; Published: Apr 15, 2003

Abstract

Every binary word with at least four letters contains a square. A. Fraenkel and J. Simpson showed that three *distinct squares* are necessary and sufficient to construct an infinite binary word. We study the following complementary question: how many *square occurrences* must a binary word contain? We show that this quantity is, in the limit, a constant fraction of the word length, and prove that this constant is 0.55080....

1 Introduction

Infinite words avoiding repetitions is a classical area in word combinatorics [2]. A famous result of A.Thue [9, 10] (see also [1]) is that squares (subwords of the form uu for a non-empty u) can be avoided on a ternary alphabet and cubes (subwords uuu) on a binary alphabet.

Different generalizations of the Thue results have been studied recently. One direction is related to considering fractional exponents. Thue showed that on the binary alphabet, a *strongly cube-free* infinite word can be constructed, i.e. a word that does not contain a subword uua, where a is the first letter of u. Putting this result in terms of fractional exponents, there exists an infinite binary word that does not contain a subword of exponent $2+\varepsilon$ for any $\varepsilon > 0$. 2 is trivially a tight bound as any binary word longer than three letters contains a square.

Generalizing this to the ternary alphabet, F. Dejean [4] showed that any exponent bigger than $7/4$ can be avoided using three letters, and this bound is tight. These results have been generalized to larger alphabets and on the other hand, to the *abelian case*

*LORIA/INRIA-Lorraine, 615, rue du Jardin Botanique B.P. 101, 54602 Villers-lès-Nancy France, Gregory.Kucherov@loria.fr

†Laboratoire Bordelais de Recherche en Informatique, 351, cours de la Libration 33405 Talence Cedex, France, ochem@labri.fr

‡Université de Metz, Laboratoire d'Informatique Théorique et Appliquée, 57045 Metz Cedex 01, France, rao@sciences.univ-metz.fr

where squares and cubes are considered modulo letter commutations. We refer to [2] for a survey of these results.

Another direction is to study limit properties of infinite words avoiding a given exponent. The following question has been studied in [7]: on the binary alphabet, what is the minimal limit fraction of one of the two letters, which allows to construct an infinite word avoiding subwords of exponent e ($e > 2$)? As an example, it has been shown that this fraction is $1/2$ for $2 < e \leq 7/3$ and strictly smaller than $1/2$ for $e = 7/3+\varepsilon$, for any $\varepsilon > 0$. For the ternary alphabet, Yu. Tarannikov [8] showed that the minimal fraction of one letter in ternary square-free words is in the interval $[1780/6481, 64/233] = [0.27464..., 0.27467...]$.

In [5], the following question has been studied: as squares cannot be avoided on the binary alphabet, how many *distinct* squares does one need in order to construct an infinite word? *Distinct* here means syntactically different squares, in contrast to *occurrences* of (possibly identical) squares. It has been proved in [5] that three distinct squares are sufficient (and necessary) to construct an infinite binary word, those squares could be 00, 11 and 0101.

In [6] was raised the complementary question of the *maximal* number of distinct squares in a binary word. It was shown that this number is linearly bounded on n (word length). More precisely, this number is always less than $2n$ and is $(n - o(n))$ for infinitely many n.

In this paper, we study the following natural question left open by [5, 6]: what is the *minimal* limit proportion of *square occurrences* in an infinite binary word? We prove that this limit exists, and prove an estimate of it, up to several decimal digits.

2 Basic definitions

Unless otherwise stated, we consider the binary alphabet $A = \{0, 1\}$. By an *infinite word* we will mean a *one-way* infinite word, also called ω-word, defined as a mapping $\mathbb{N} \to A$. The set of infinite words over A is denoted A^ω.

A *square* (in a word) is a subword uu, where u is a non-empty word. For a word $w \in \{0,1\}^*$, let $s(w)$ be the number of (possibly overlapping) square occurrences in w. For $n \in \mathbb{N}$, define $m(n) = \min_{|w|=n} s(w)$. For example, $m(3) = 0$, $m(4) = 1$, $m(5) = 1$, $m(6) = 2$. Values of $m(n)$ for low n as well as some values for big n are shown in the Appendix.

If w is a binary word, define $\overline{w}$ as the word obtained by exchanging 0 and 1 in w.

3 Limit proportion of square occurrences in binary words

The quantity we are interested in in this paper is the limit value of $\frac{m(n)}{n}$. We first show that this limit indeed exists.

Lemma 1. *For every $n \in \mathbb{N}$, for every $k \in \mathbb{N}$, $k > 1$,* $\frac{m(n)}{n} > \left(1 + \frac{n-k^2}{n(k-1)}\right) \times \frac{m(k)}{k}$.

Proof. Take a word of length n with $m(n)$ square occurrences and consider $\left\lfloor \frac{n-1}{k-1} \right\rfloor$ subwords of length k overlapping by one letter. Each subword has at least $m(k)$ square occurrences and every square occurrence is entirely contained in at most one subword. Thus, $m(n) \geq \left\lfloor \frac{n-1}{k-1} \right\rfloor \times m(k)$, which gives

$$\frac{m(n)}{n} \geq \left\lfloor \frac{n-1}{k-1} \right\rfloor / \left(\frac{n}{k}\right) \times \frac{m(k)}{k} > \left(\frac{n-1}{k-1} - 1\right) / \left(\frac{n}{k}\right) \times \frac{m(k)}{k} = \left(1 + \frac{n-k^2}{n(k-1)}\right) \times \frac{m(k)}{k}.$$

$\square$

Note that a binary word can contain $\Theta(n^2)$ square occurrences (consider the word 1^n) and as many as $\Theta(n \log n)$ occurrences of *primitive squares* (see [3]), i.e. squares uu such that u can not be written as v^k for $k \in \mathbb{N},\ k \geq 2$. The infinite word constructed in [5] contains only distinct squares 00, 11 and 0101, and therefore at most one square can start at each position of this word. This implies that $m(n) < n$.

Theorem 2. *The sequence* $\frac{m(n)}{n}$ *converges.*

Proof. Due to the remark above, $\left\{\frac{m(n)}{n}\right\}$ is bounded. Then, by the Bolzano-Weierstrass theorem, there exists an accumulation point. We deduce from Lemma 1 that $\frac{m(n)}{n} > \frac{m(k)}{k}$ for $n \geq k^2$. This proves that there is a unique accumulation point $\mathcal{M} = \lim_{n\to\infty} \frac{m(n)}{n}$. $\square$

Our goal is to estimate $\mathcal{M}$. The following lemma is useful.

Lemma 3. *For every* $k \in \mathbb{N},\ k > 1$, $\mathcal{M} \geq \frac{m(k)}{k-1}$.

Proof. Using Lemma 1, we have

$$\mathcal{M} = \lim_{n\to\infty} \frac{m(n)}{n} \geq \lim_{n\to\infty} \left(1 + \frac{n-k^2}{n(k-1)}\right) \times \frac{m(k)}{k} = \left(1 + \frac{1}{k-1}\right) \times \frac{m(k)}{k} = \frac{m(k)}{k-1}.$$

$\square$

4 Upper bound

To approach the upper bound, we first estimate the number of square occurrences in A. Fraenkel and J. Simpson's construction [5]. The infinite binary word constructed in [5] is obtained by first translating a specific square-free word W_3 over the ternary alphabet $\{a_1, a_2, a_3\}$ to a word W_5 over the quinary alphabet $\{a_1, a_2, a_3, a_4, a_5\}$, and then by applying to W_5 a morphism to the binary alphabet $\{0, 1\}$. The first step is such that the occurrences of a_1, a_2, a_3 in W_3 and W_5 are in bijective correspondence, and for each occurrence of a_3 in W_3, an occurrence of either a_4 or a_5 is introduced in W_5. Assume that $\mathcal{X}$ is the limit fraction of a_3 in the initial ternary word W_3. Let $\mathcal{P}_{a_1a_2a_3}$ (respectively $\mathcal{P}_{a_4a_5}$) be the limit proportion of letters a_1, a_2, a_3 (respectively a_4, a_5), counted together, in word W_5. According to the above description of W_5, $\mathcal{P}_{a_1a_2a_3} = 1/(1+\mathcal{X})$ and $\mathcal{P}_{a_4a_5} = \mathcal{X}/(1+\mathcal{X})$.

At the second step, the morphism maps each of $\{a_1, a_2, a_3\}$ to a binary word of length 12, and each of $\{a_4, a_5\}$ to a binary word of length 14. Moreover, each image of $\{a_1, a_2, a_3\}$ adds 7 square occurrences to the resulting binary word (6 squares inside the image and one across the border with the previous image), and each image of $\{a_4, a_5\}$ adds 8 of those (7 and 1 respectively). We conclude that the limit proportion of square occurrences in the final binary word of [5] is

$$\frac{7 \cdot \mathcal{P}_{a_1a_2a_3} + 8 \cdot \mathcal{P}_{a_4a_5}}{12 \cdot \mathcal{P}_{a_1a_2a_3} + 14 \cdot \mathcal{P}_{a_4a_5}} = \frac{7 + 8 \cdot \mathcal{X}}{12 + 14 \cdot \mathcal{X}}.$$

On the other hand, $\mathcal{X}$ can be bounded by $1/4 \leq \mathcal{X} \leq 1/2$, as there must be at least one a_3 in every subword of length 4 and at most one a_3 in every subword of length two[1]. Therefore, the proportion of square occurrences in the word of [5] is between $11/19 = 0.5789..$ and $18/31 = 0,5806....$ We now show that the minimal proportion $\mathcal{M}$ is smaller than that, by showing a smaller upper bound.

Our construction is based on the following pattern of length 187, noticed when computing long words that realize the minimal number of square occurrences for their length[2].

$$\begin{aligned} w = \; & 0100110100011001011000110100110001011001010011010001100101110 \\ & 0110100111001011001110100110101100101110011010 0X1100101100Y1 \\ & 1010011000101100101001101000110010110001101001100010110011101 \\ & 00110 \end{aligned}$$

The following words are obtained by substituting in different ways variables X and Y in w and then by concatenating the resulting words with their complements.

$$\begin{aligned} v_a &= w|_{X\to 0, Y\to 0} \\ v_b &= w|_{X\to 1, Y\to 0} \\ v_c &= w|_{X\to 1, Y\to 1} \end{aligned}$$

$$\begin{aligned} w_a &= v_a\overline{v_a}, \\ w_b &= v_a\overline{v_b}, \\ w_c &= v_b\overline{v_c} \end{aligned}$$

w_a, w_b and w_c are of size 374 and a computer check shows that each of them has 204 square occurrences.

Consider the morphism h defined by $h(a) = w_a$, $h(b) = w_b$ and $h(c) = w_c$. Let $t \in \{a, b, c\}^*$ be a square-free ternary word. Then $h(t)$ is a word of size $374 \times |t|$.

[1]These bounds are not the best possible but are sufficient for our purpose here. The lower bound can be made better using the result of [8] (see Introduction). Note further that [5] uses the subclass of ternary square-free words which avoid $a_1a_3a_1$ and $a_2a_3a_2$. This puts strong additional constrains on $\mathcal{X}$.

[2]We will describe in the end of Section 5 how long words realizing the minimal number of squares have been computed.

Concatenating two different words of $\{w_a, w_b, w_c\}$ creates two new squares crossing the boundary: 0101 and 1010. We show that there are no other new squares in $h(t)$.

Lemma 4. *Each square occurrence of $h(t)$ either is located inside the image of a letter of t, or is one of the squares* 0101 *and* 1010 *crossing the boundary between two adjacent letter images. Consequently, $h(t)$ contains* $(206 \times |t| - 2)$ *square occurrences.*

Proof. Assume that $h(t)$ contains a square, of size k, which is neither of those specified in the lemma.

If $k < 4 \times 374$, this square is contained in the image by h of a subword of t of length at most 5. However, a computer check shows that for every ternary square-free word t' of size at most 5, $h(t')$ contains only the squares specified in the lemma.

Assume $k \geq 4 \times 374$ and let uu be the square under consideration. Since $|u| \geq 2 \times 374$, one of the words $\{w_a, w_b, w_c\}$ is a subword of u, and therefore has two occurrences in $h(t)$ at distance $|u|$. This word must be a subword of a word $h(xy) = w_x w_y$, $x, y \in \{a, b, c\}$. A direct verification shows that any word of $\{w_a, w_b, w_c\}$ can occur in a word $w_x w_y$, $x, y \in \{a, b, c\}$, only as a suffix or as a prefix but not as a proper subword. This implies that $|u|$ is a multiple of 374, and k is a multiple of 2×374. Furthermore, this square cannot be centered at the boundary of two letter images, as this would imply that uu is the image of a subword of t and this subword is a square too (note that the inverse image of h is unique), which would contradict to the square-freeness of t.

Now, we note that the minimal subword of t such that $h(t)$ contains a square of size $2 \times 374 \times l$ must be of the form $\alpha v \beta v \gamma$, where α, β, γ are letters, and v is a word of size $l - 1$ (see Figure 1).

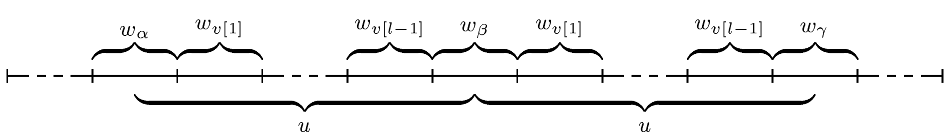

Figure 1: Square uu occurring in $h(t)$

w_a, w_b and w_c differ only in 3 positions: positions 109, 296 and 307. The letters at those positions are respectively 0 1 1 for w_a, 0 0 1 for w_b and 1 0 0 for w_c.

If the center of the square is before position 296, the letters at positions 296 and 307 are the same in w_β and in w_γ. This implies that $w_\beta = w_\gamma$, thus $\beta = \gamma$ and therefore t contains the square $v\gamma v\gamma$. By a similar argument, if the center of the square is after position 296, then t contains the square $\alpha v \alpha v$. In either case, this contradicts our assumption that t is square-free. We conclude that $h(t)$ does not contain squares other than those specified in the lemma, and then $h(t)$ has $(206 \times |t| - 2)$ square occurrences. □

Corollary 5. $\mathcal{M} \leq \frac{103}{187} = 0.55080\ldots$

5 Lower bound

In [8], Yu. Tarannikov introduced a method for obtaining lower bounds, that can be applied to our case. We summarize it in the following lemma. Recall that $s(w)$ is the number of square occurrences in w.

Lemma 6. *For $\xi \in \mathbb{R}$, define*

$$A(\xi) = \left\{ w \in \{0,1\}^* |\ \textit{for every prefix } w[1..k] \textit{ of } w,\ \frac{s(w[1..k])}{k} \le \xi \right\}.$$

Then

(i) $A(\xi)$ is finite iff $\xi < \mathcal{M}$,

(ii) there exists $w \in \{0,1\}^\omega$ such that for every finite prefix $w[1..k]$, $\frac{s(w[1..k])}{k} \le \mathcal{M}$.

Proof. Direct application of the corresponding proofs of [8]. □

According to condition (i) above, if for some ξ, $A(\xi)$ is shown to be finite, then ξ is a lower bound for $\mathcal{M}$. The method consists then in exploring $A(\xi)$ and showing that it is "saturated" at a certain word length and cannot be extended for longer words.

The interest of exploring $A(\xi)$ is that its definition may allow to reduce the search space. In our case however, we were unable to obtain a good lower bound by a direct application of Lemma 6, as the search space quickly became prohibitively big. To obtain a good lower bound, we use the following extension of Lemma 6. For a word $u \in \{0,1\}^*$, let

$$A_u(\xi) = \left\{ w \in \{0,1\}^* |\ \text{for every prefix } w[1..k] \text{ of } w,\ \frac{s(uw[1..k]) - s(u)}{k} \le \xi \right\}.$$

Lemma 7. *Fix $r \in \mathbb{N}$. If for some ξ and all $u \in \{0,1\}^r$, $A_u(\xi)$ is finite, then $\xi < \mathcal{M}$.*

Proof. Let $l = \max_{w \in A_u(\xi)} |w|$. There exists $\varepsilon > 0$ such that $\forall u \in \{0,1\}^r$, $\forall w \in \{0,1\}^*, \exists k \le l+1$ such that $\frac{s(uw[1..k]) - s(u)}{k} \ge \xi + \varepsilon$.

Let w be a binary word of size $n > r$ such that $s(w) = m(n)$. Let $k_0 = r$. For $i \ge 1$, let $v_{i-1} = w[k_{i-1} - r + 1..k_{i-1}]$ and let k_i be the smallest position, if exists, such that

$$\frac{s(v_{i-1}w[k_{i-1}+1..k_i]) - s(v_{i-1})}{k_i - k_{i-1}} \ge \xi + \varepsilon.$$

By the above remark, $k_i - k_{i-1} \le l+1$. Let q be the last i for which k_i has been defined. Then $k_q \ge n - l$. We then have

$$\begin{aligned} m(n) = s(w) &\ge \sum_{i=1}^{q} \left(s(v_{i-1}w[k_{i-1}+1..k_i]) - s(v_{i-1}) \right) \\ &\ge (\xi+\varepsilon) \sum_{i=1}^{q} (k_i - k_{i-1}) = (\xi+\varepsilon)(k_q - r) \ge (\xi+\varepsilon)(n-l-r). \end{aligned}$$

Then $\mathcal{M} = \lim_{n\to\infty} \frac{m(n)}{n} \ge \xi + \varepsilon$. □

Similar to Lemma 6, applying Lemma 7 consists in exploring, for some r and ξ, the sets $A_u(\xi)$ for all u with $|u| = r$. However, Lemma 7 allows to reduce substantially the search space, in comparison to Lemma 6. Thus, showing that $\xi = 0.55$ is a lower bound took several hours for $r = 1$, several seconds for $r = 2$, and a fraction of second for $r = 3$. For $r = 3$, we managed to show that $A_u(0.5508)$ is finite for all u, $|u| = 3$. The verification has taken more than 19 hours of CPU time on an AMD Athlon™ 1.4 GHz computer. $A_u(0.5508)$ reaches its biggest size for $u = 000$ and $u = 111$, with longest word length 5195.

Together with Corollary 5, we obtain

Theorem 8. $\mathcal{M} = 0.55080....$

Finally, we were also able to compute $m(n)$ for all n up to about 3300 using an optimized search for words realizing the minimal number of squares. Below we briefly describe the method used for this computation.

For a word u and for $n \geq |u|$, let $m(u, n)$ be the smallest number of square occurrences in a word of length n with prefix u. Then for all $p \geq 0$ and $n \geq p$, $m(n) = \min_{u \in \{0,1\}^p}\{m(u, n)\}$.

To compute $m(n)$, together with a witness word, we first fix some p (6 in our case). For every word u of size p and for every $n \geq p$, we compute and store $m(u, n)$. We proceed successively for $n \geq p$ and for each n, we start by computing an upper bound B on $m(u, n)$ from the witness word for $m(u, n-1)$, by appending 0 or 1 to it.

We then try to construct a word $w[1..n]$ containing at most $B - 1$ squares. For each prefix $w[1..k]$ of such a word, we must have

$$s(w[1..k]) < B - m(w[k-p+1..k], n-k+p) + s(w[k-p+1..k]),$$

since we know that $w[k+1..n]$ must add to $w[1..k]$ at least $(m(w[k-p+1..k], n-k+p) - s(w[k-p+1..k]))$ squares. Thus, we explore the tree of all words w of length at most n satisfying the above inequality for every prefix $w[1..k]$. If it is not verified, we "cut the branch". If we succeed to construct a word $w[1..n]$ such that each its prefix verifies the above inequality it implies that we came up with a smaller upper bound on $m(u, n)$ and a new witness word. We use this new upper bound in the further search. At the end of the search, we obtain the minimum value of $m(u, n)$ and a corresponding witness word.

This method allowed us to reduce the search space drastically and to compute $m(n)$ for n as big as 3300. Some selected values of $m(n)$ are given in Appendix. These data can be also used to obtain lower bounds on $\mathcal{M}$ as implied by Lemma 3 for example. In this way, the best lower bound results from $m(3298) = 1815$ and is then $1815/3297 = 0.5505...$, which is still smaller than the one we were able obtain using Lemma 7.

6 Weaker lower bounds

In this section we show another way to obtain lower bounds that are in general weaker than those obtained by the methods of the previous section. However, the construction is interesting on its own, and comes through weakening the definition of $\mathcal{M}$.

For $k \in \mathbb{N}^*$, let $s_k(w)$ be the number of square occurrences of size at most $2k$ in the binary word w. For $n \in \mathbb{N}$, we define $m_k(n) = \min_{|w|=n} s_k(w)$. For the same reason as for $m(n)$, for every $k \in \mathbb{N}$, the sequence $\frac{m_k(n)}{n}$ converges as $n \to \infty$. Let $\mathcal{M}_k$ be its limit. Note that $\{\mathcal{M}_k\}$ is an increasing sequence bounded by $\mathcal{M}$.

Assume that w is a word such that $|w| > k$ and $s_k(ww) = 2s_k(w)$. Then, for all $q \in \mathbb{N}$, $s_k(w^q) = qs_k(w)$, since no square of length at most $2k$ can span over more than two occurrences of w. We then have $\mathcal{M}_k \leq \frac{s_k(w)}{|w|}$. Using this argument, we can compute an upper bound on $\mathcal{M}_k$ by finding an appropriate word w. Specifically, the words 01, 001, 001011, 1100101100011010011000101100101001101000 and 00010110011101001101011001011100110100111001011001110100110101100101110011010 0011001011000110100110001011001010011010001100101100011010011 prove respectively that $\mathcal{M}_1 = 0$, $\mathcal{M}_2 \leq \frac{1}{3}$, $\mathcal{M}_5 \leq \frac{1}{2}$, $\mathcal{M}_{39} \leq \frac{11}{20}$ and $\mathcal{M}_{137} \leq \frac{38}{69}$. These words have been found by a computer search using a method similar to the one described at the end of Section 5.

We now introduce a method for computing exact values of $\mathcal{M}_k$. A *weighted directed graph* $G = (V, A)$ is a directed graph with a weight function on arcs $\rho : A \to \mathbb{N}$. A *pass* is a finite sequence $P = v_1 v_2 \dots v_k$ of vertices of G, such that for every $i \in \{1, 2, \dots, k-1\}$, $< v_i, v_{i+1} > \in A$. A *cycle* is a pass $C = v_1 v_2 \dots v_k v_1$. C is a *simple cycle* if for $i, j \in \{1, 2, \dots, k\}$ $v_i \neq v_j$ provided $i \neq j$. The *size* of a pass $P = v_1 v_2 \dots v_k$, denoted $|P|$, is $k - 1$. In particular, a cycle $C = v_1 v_2 \dots v_k v_1$ has size k. The *weight* of a pass $P = v_1 v_2 \dots v_k$, denoted $\rho(P)$, is $\sum_{i=1}^{k-1} \rho(< v_i, v_{i+1} >)$.

Let us fix $k \in \mathbb{N}^*$ and consider the weighted directed graph G_k in which the vertices are binary words of size $(2k-1)$, and each vertex has two outgoing arcs, one for 0 and one for 1. For a vertex corresponding to a word v and an outgoing arc a ($a \in \{0, 1\}$), the destination of the arc is the vertex corresponding to the word $va[2..2k]$ (i.e. the word va without the first letter). The weight of this arc is the number of squares that are suffixes of va. Note that G_k has 2^{2k-1} vertices.

Lemma 9. *Let* $\mathcal{M}'_k = \min_{C \textit{ simple cycle of } G_k} \frac{\rho(C)}{|C|}$. *Then* $\mathcal{M}_k = \mathcal{M}'_k$.

Proof. We first note that if C is a (not necessarily simple) cycle in G_k, then $\rho(C) \geq \mathcal{M}'_k \cdot |C|$. This can be seen by naturally decomposing C into a parenthesis-like structure of simple cycles. Each arc of C belongs to exactly one of those simple cycles. This implies that $\rho(C)$ equals the sum of weights of all those cycles, each of which is at least $\mathcal{M}'_k$ related to its length.

Now let t be a word of size $n \geq 2k-1$ such that $s_k(t) = m_k(n)$ and let P_t be the pass in G_k corresponding to t whose source vertex corresponds to the prefix $t[1..2k-1]$ ("spelling pass"). Note that $|t| = |P_t| + 2k - 1$. We decompose P_t through the following iterative procedure. Find the first vertex in the pass which occurs at least twice, and consider the cycle between its first and last occurrence. Then iterate the procedure on the remaining part of the pass. As a result, we obtain a decomposition $P_t = p_0 C_1 p_1 C_2 \dots p_{q-1} C_q p_q$, where C_i are cycles and p_j are passes without cycle. Note that every vertex appears in at most one of these passes, therefore $\sum_{i=0}^{q} |p_i| \leq 2^{2k-1}$.

We then have

$$m_k(n) \geq \sum_{i=1}^{q} \rho(C_i) \geq \mathcal{M}'_k \sum_{i=1}^{q} |C_i| = \mathcal{M}'_k \left(n - (2k-1) - \sum_{i=0}^{q} |p_i| \right) \geq \mathcal{M}'_k (n-(2k-1)-2^{2k-1}).$$

Therefore,

$$\mathcal{M}_k = \lim_{n\to\infty} \frac{m_k(n)}{n} \geq \mathcal{M}'_k.$$

To show the inverse inequality, we choose a simple cycle C_{min} such that $\mathcal{M}'_k = \frac{\rho(C_{min})}{|C_{min}|}$. Consider words t_q, defined for $q > 0$ by $P_{t_q} = (C_{min})^q$. We then obtain

$$\mathcal{M}_k \leq \lim_{q\to\infty} \frac{s_k(t_q)}{|t_q|} \leq \lim_{q\to\infty} \frac{q \cdot \rho(C_{min}) + k^2 - k}{q|C_{min}| + 2k - 1} = \mathcal{M}'_k.$$

□

Figure 2 shows the graph G_2, and a simple cycle C realizing the minimal ratio $\mathcal{M}'_2 = \frac{w(C)}{|C|} = \frac{1}{3}$.

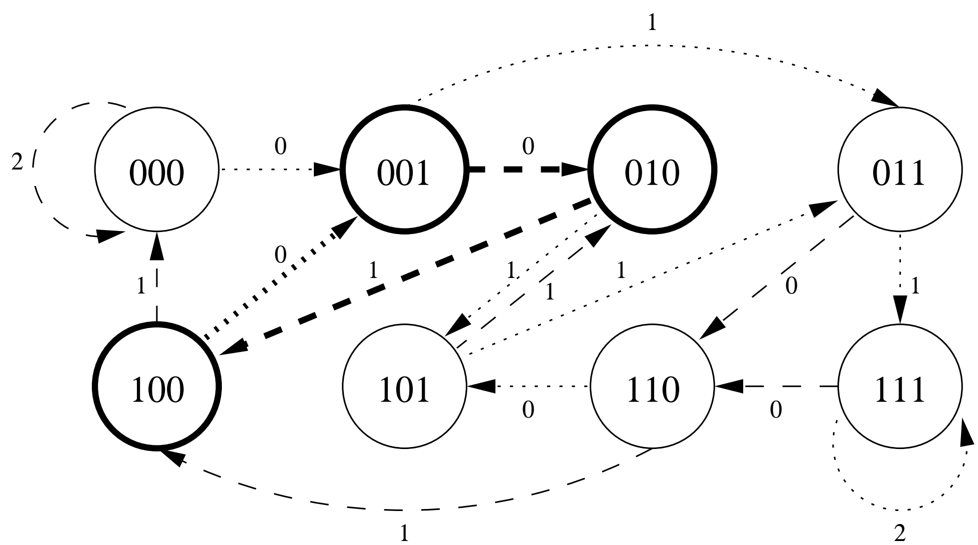

Figure 2: The graph G_2 with an optimal simple cycle (in bold). Dashed arcs are 0-arcs and dotted arcs are 1-arcs. The numerical label of the arc is its weight.

Lemma 9 allows us to compute $\mathcal{M}_k$ for small k. Using the computer, we obtained in particular the values $\mathcal{M}_2 = \frac{1}{3}$, $\mathcal{M}_3 = \frac{1}{2}$ and $\mathcal{M}_6 = \frac{11}{20}$.

Using the upper bounds obtained in the beginning of this section and the fact that $\{\mathcal{M}_k\}$ is non-decreasing, we have

$$\mathcal{M}_1 = 0,\ \mathcal{M}_2 = \frac{1}{3},\ \mathcal{M}_3 = \mathcal{M}_4 = \mathcal{M}_5 = \frac{1}{2},\ \mathcal{M}_6 = ... = \mathcal{M}_{39} = \frac{11}{20}.$$

This shows once again that $\mathcal{M} \geq \frac{11}{20} = 0.55$.

7 Conclusions

Mysterious constants abound in word combinatorics (see [2]). The exact values of most of them is not known and in most cases only estimations, more or less precise, are available. In this paper we introduced and studied a new remarkable constant – the limit minimal fraction of the number of square occurrences in binary words. We were able to obtain a very good estimation of this constant (0.55080...) but its exact value remains unknown.

An interesting question is to study *which* squares are needed to realize an infinite word with minimal number of squares. We conjecture that squares of length 2 or 4 are sufficient, as is the case in our construction from Section 4.

References

[1] J. Berstel. Axel Thue's work on repetitions in words. Invited Lecture at the 4th Conference on Formal Power Series and Algebraic Combinatorics, Montreal, 1992, June 1992. disponible à l'adresse `http://www-igm.univ-mlv.fr/~berstel/index.html`.

[2] C. Choffrut and J. Karhumäki. Combinatorics of words. In G. Rozenberg and A. Salomaa, editors, *Handbook on Formal Languages*, volume I, pages 329–438. Springer Verlag, Berlin-Heidelberg-New York, 1997.

[3] M. Crochemore. An optimal algorithm for computing the repetitions in a word. *Information Processing Letters*, 12:244–250, 1981.

[4] F. Dejean. Sur un théorème de Thue. *J. Combinatorial Th. (A)*, 13:90–99, 1972.

[5] A. Fraenkel and J. Simpson. How many squares must a binary sequence contain? *Electronic Journal of Combinatorics*, 2(R2):9pp, 1995. `http://www.combinatorics.org/Journal/journalhome.html`.

[6] A. Fraenkel and J. Simpson. How many squares can a string contain? *J. Combinatorial Theory (Ser. A)*, 82:112–120, 1998.

[7] R. Kolpakov, G. Kucherov, and Y. Tarannikov. On repetition-free binary words of minimal density. *Theoretical Computer Science*, 218(1), 1999.

[8] Y. Tarannikov. The minimal density of a letter in an infinite ternary square-free word is 0.2746.... *Journal of Integer Sequences*, 5(2):Article 02.2.2, 2002. `http://www.math.uwaterloo.ca/JIS/`.

[9] A. Thue. Über unendliche Zeichenreihen. *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania*, 7:1–22, 1906.

[10] A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske Vid. Selsk. Skr. I. Mat. Nat. Kl. Christiania*, 10:1–67, 1912.

Appendix

n	$m(n)$	witness word
1	0	0
2	0	01
3	0	010
4	1	0100
5	1	01001
6	2	010010
7	2	0100110
8	2	01001101
9	3	010011010
10	4	0100110100
11	4	01001101001
12	5	010011010010
13	5	0100110100010
14	6	01001101000101
15	6	010011010001101
16	7	0101100101110010
17	7	01001101000110010
18	8	010011010001100101
19	8	0100110100010110010
20	9	01001101000101100101
21	10	010011010001011001010
22	10	0100110100010111001101
23	11	01001101000101110011010
24	11	010011010001011101001101
25	12	0100110100010111010011010
26	12	01001101000101100101001101
27	13	010011010001011001010011010
28	13	0100110100010110011101001101
29	14	01001101000101100111010011010
30	15	010011010001011001110100110101
31	15	0100110100010111010011010110010
32	16	01001101000101110100110101100101
33	16	010011010001011101001100010110010
34	17	0100110100010111010011000101100101
35	17	01001101000101110010110011101001101
36	18	010011010001011100101100111010011010
37	18	0100110100010110011101001100010110010
38	19	01001101000101100111010011000101100101
39	20	010011010001011001110100110001011001010
40	20	0100110100010111010001101001100010110010

Table 1: First 40 values of $m(n)$ together with a witness word

n	50	100	500	1000	1500	2000	2500	3000	3298	3300
$m(n)$	25	53	273	549	824	1099	1375	1650	1815	1816

Table 2: Values of $m(n)$ for selected big n